Data Cleaning & Summarising

PRACTICAL DATA SCIENCE WITH PYTHON – ASSIGNMENT 1

Student name: Hue Phuong Le
Student ID: s3687477

Data Preparation

My explanation of how I addressed the task 1, including the steps of dealing with the potential issues/errors, and I will create one sub-section for each type of errors.

Data Entry Error

With frequency table, it is detected that Pos column has three data entry errors which are ‘PFA’, ‘SF.’ and ‘SGA’. Moreover, the specification listed out a set of values that they can be. Therefore, I easily fixed them with simple assignment statements and if-then-else rules. They were corrected to ‘PF’, ‘SF’ and ‘SG’ respectively.

Redundant Whitespace and Capital Letter Mismatches

There are two columns that face this type of errors which are Pos and Tm column. These errors were simply fixed by removing both leading and trailing spaces with an inbuilt function and then replacing all strings with their uppercase version.

Impossible Values and Sanity Checks

The frequency table is used to check for impossible values from Age column, and I figured out there are two impossible values which are ‘-19’ and ‘280’. Because other corresponding values from these rows are normal, what’s more, a number of players has the age of either 19 or 28. Therefore, I they should be data entry errors and were replaced with 19 and 28 respectively.

Based on the specification, some sanity checks are performed for some columns below:

- All values in columns with data type ‘int64’ have to be greater than or equal to 0.
- Values from G are smaller than or equal to 82 and cannot be smaller than values from GS.
- Values from FGA cannot be smaller than FG. Similarly for 3-Point, 2-Point and Free Throws.
- Values from PTS must be less than 2000. Some errors appeared and the specification already provided a formula how to calculate this column, I safely re-calculated them. Thus, I do not have to remove and lose these records.
- Values from PF cannot be greater than six times values from G.

Except for the values from PTS column, all the sanity checks did not give me any unexpected errors.

Missing Values

Null values come from three columns which are 3P%, 2P% and FT%. Technically, they are calculated by a division of values from other columns, and if the denominator is 0, it is invalid, leading to null.

Nonetheless, it is complicated if I fix the values. For example, in 3P, 3PA and 3P% columns; in order to fill in the missing values in 3P%, 0 values from 3PA must be replaced by any values except 0. And 3PA cannot be less than 3P, hence I cannot just substitute automatically with the mean/median value. Moreover, there are 61 null values, replacing all of them would have impact on my descriptive statistics. As a result, in this case, I removed all of these records because compared to the total records I have, it does not affect that much.

Invalid Values Due to Calculation

I decided to re-calculate values from columns FG%, 3P%, 2P%, FT%, TRB to avoid any errors due to incorrect calculation.

Outliers

Drawing historgram is the most appropriate approach to find outliers. I found outliers from the multiple columns below and treated them differently:

- There are two outliers from FG%, 3PA, one from 3P, four from 2P%, FT and five from FT%. They are completely deleted for some reasons. Firstly, there are other inappropriate values from other features in these records. Secondly, the number of records is not that much compared to the number of records in my dataset, and they do not represent anything. Thirdly, the outliers from 3PA are greater than 350 (much higher compared to the nearest one which is 308). Lastly, these values depend on or affects values in other columns.
- For 3P%, initially there is an outlier with much higher value, the relating record is removed because the value depends on values from 3P and 3PA. After that, by drawing the graph for double check, there are nearly 40 more outliers. They were kept due to high frequency.
- There are five outliers from BLK column and one from PF. Because they do not depend on or affect any other values, I replaced them with the mean value to not lose data.

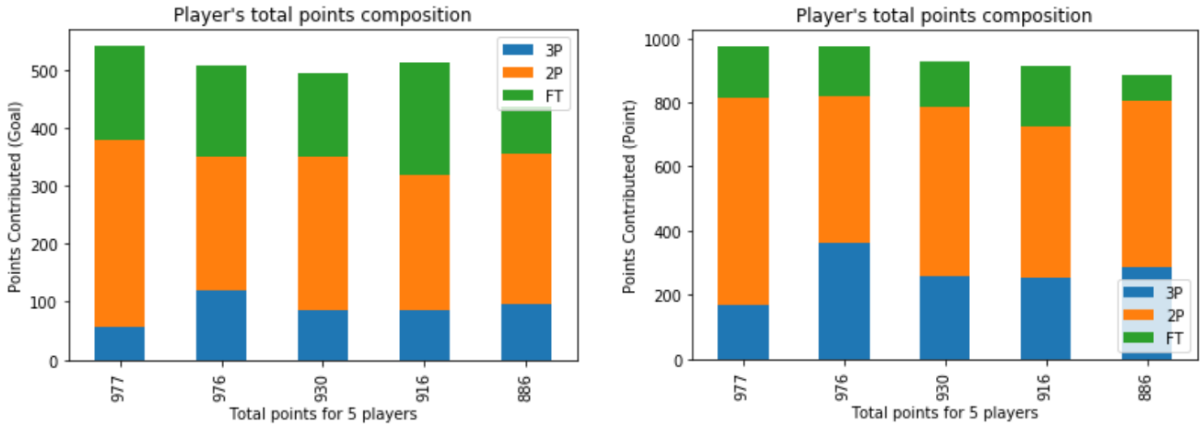
Special Characters Due to Encoding

There are some strange characters appeared for player's name due to encoding from excel file to the dataframe. Nevertheless, to keep simplify, I remain them the same because the player's name is not important for further investigation.

Data Exploration

Task 2.1

Below are the graphs of player's total points composition (the left graph is the number of 3-Point Field (3P), 2-Point Field (2P) and Free Throw (FT) goals & the right one is the number of points from 3P, 2P and FT contributes to the total points)



It is clearly showed that for top five players, their total points and goals are almost contributed by the 2P. For their number of goals, the second factor contributed is 3P and then FT. However, because players get three points for each 3P goal and only one for FT, 3P contributes more to the total points compared to FT.

Specifically, player with the highest total points has the highest number of not only goals (sum of 3P, 2P and FT) but also 2P goals. It is analyzed that 66% of his total points come from 2P, and 17% from either 3P or FT.

Moreover, for the second and fourth players, although second player has less not only number of goals than fourth player (506 smaller than 512) but also amount of 2P points (460 smaller than 468)

and FT (156 smaller than 193), the second player still has more total points due to 3P. The second player has 120 goals from 3P while the fourth player only has 85 goals. Specifically, for the second player, the number of 3P only takes 24% in the total number of goals but it contributes 37% to the total points.

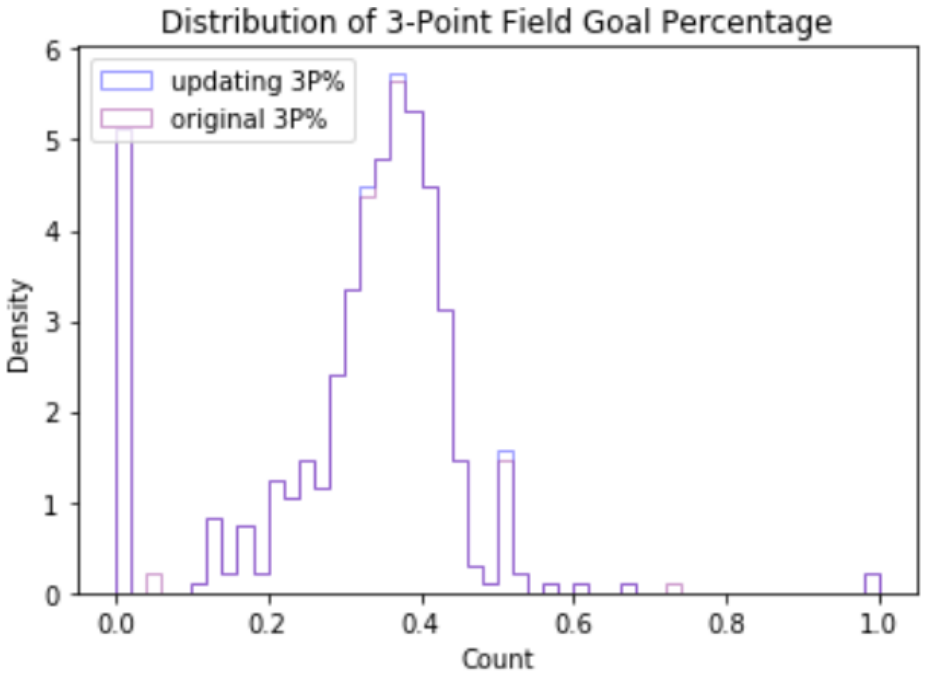
Another comparison is for the third and last players, their total points consist of a similar total amount of points from 3P and 2P (786 and 806 for the third and last players respectively). However, it still leads to a different ranking due to the points from FT. The third player got nearly double points from FT compared to the last player (144 compared to 80).

Lastly but most importantly, except for the first players, around half of their total points come from 2P (between 47% and 58%), and then from 3P (from 28% to 37%), lastly is FT (only from 9% to 21%).

Task 2.2

From the data cleaning, I already ensured values for these columns cannot be negative. And when comparing 3P and 3PA, it is clearly showed that they are all valid. The records with value of 0 from 3PA are also removed. In case values from 3P and 3PA are exceptionally high or low, we cannot consider them entry error because they are still reasonable.

Consequently, I believe the errors might come from calculating or inputting 3P% incorrectly. By visualizing values from both original and updating dataset in one graph, the result is showed as below:



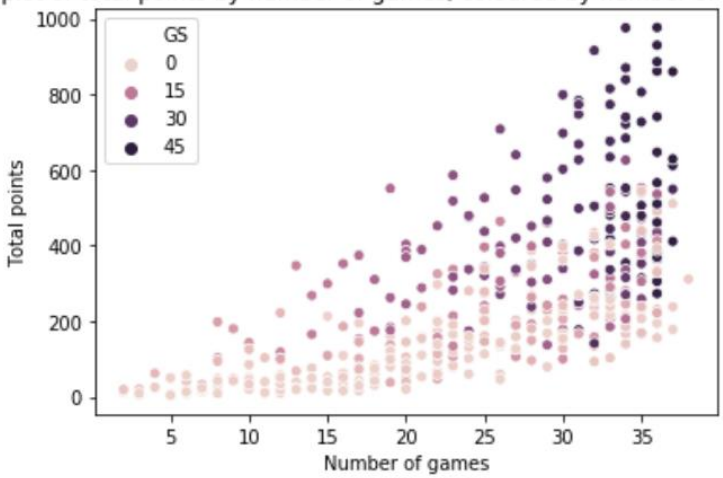
Obviously, it is clearly seen that after re-calculating 3P%, the line for updating 3P% (blue color) does not match the line for original 3P% (purple color).

By using inbuilt function to compare values from 3P% column between two datasets (original and updating), there are a total of 76 errors. They can be simply fixed by re-calculate the values for 3P% (which I already did from data cleaning part).

Task 2.3

Figure 1: Relationship between player's total points, games & games started

Figure 4: Scatterplot of total points by number of games, coloured by number of games played as starter

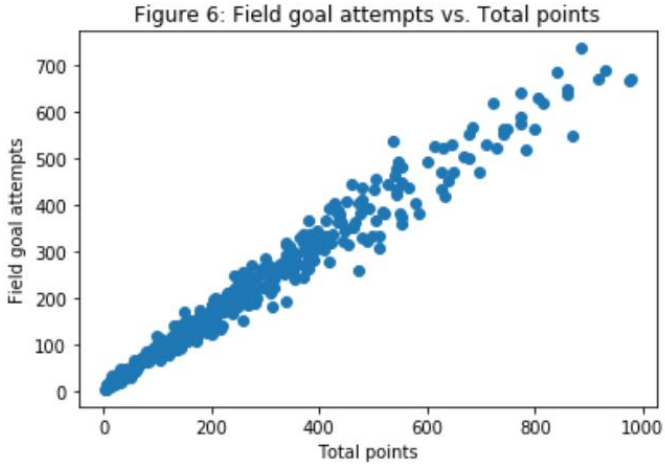


The scatterplot in the figure above illustrates apparent distinction between player's total points and number of games, grouped by the number of games played as starter.

It is clearly seen that for players which attend from 0 to 30 games as starter, the more games they play, the more total points they get. However, it is predictable that their points only experience a slight increase towards 500 points.

By contrast, for players with less than 45 starter games, they all attend around 30 to 35 games, and their total points are higher and more varied (200-600 points), which is unpredictable. However, the total points are higher and more stable as the data funnel towards 1000 points.

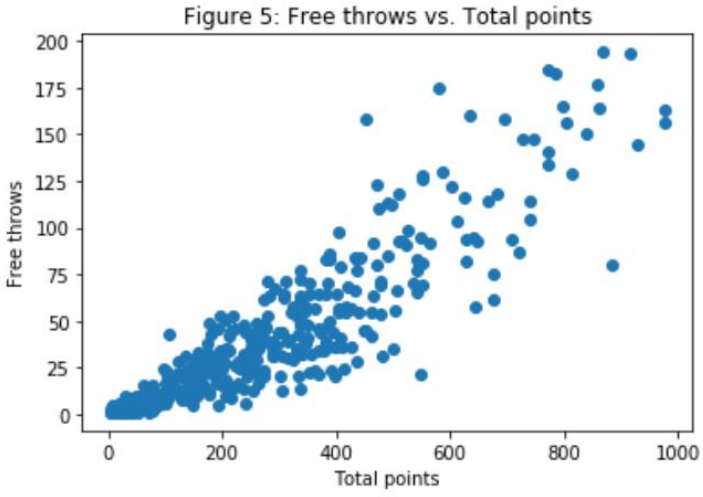
Figure 2: Relationship between player's total points and field goal attempts



There is a strong correlation between player's total points and field goal attempts. This graph shows a steady increase (total points increase from 0 to nearly 1000 points while field goal attempts increase from 0 to around 700 attempts).

Moreover, obviously there is a positive linear relationship with two of them as well. The more field goal attempts players reach, the higher total points they collect. Upon observation, it is predictable that the field goal attempts including 2P and 3P have a tendency to decide whether player ends up with low or high total points.

Figure 3: Relationship between player's total points and free throws



Like figure 2, this graph also shows a positive linear relationship between player's total points and free throws. Differently, there is a slight increase (0 to nearly 200 for free throws and 0 to nearly 1000 for player's total points), which indicates that the relationship between total points and field goal attempts is stronger than those with free throws.